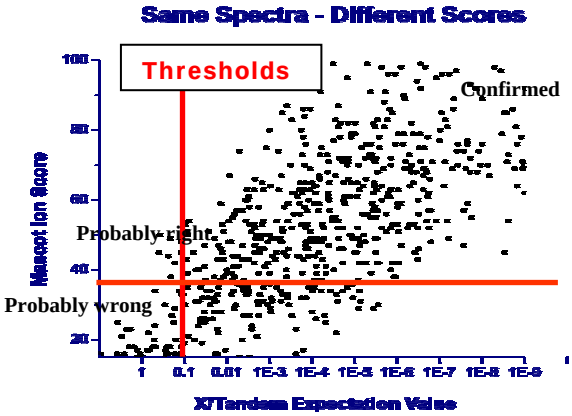
Innovations in Proteomics: Reference Library of Peptide Ion Fragmentation Spectra for MS Users

A library of peptide ion fragmentation spectra is being developed and tested, to enable discovery scientists to more rapidly and reliably identify peptides and proteins. The library is based on both a 'Consensus spectrum' and the best experimental spectra from all spectra matching a given peptide ion. All spectra in this library originated from peptide ions generated by electrospray ionization in LC-MS/MS experiments. The library is being incorporated into several proteomics 'pipelines' for integration into practical proteomics analysis.

S.E. Stein, Y. Mirokhin, P. Neta, Q.-L. Pu, J. Roth, P.A. Rudnick, M.A. Wingerd, X. Yang L. Kilpatrick (Div. 838)



Mass spectrometric (MS) measurements on proteins and peptides have attracted increasing interest over the last decade as the field of proteomics became dominant in many biochemical and biological investigations. However, current methods for peptide identification use only a fraction of the information in the mass spectra of peptide ions. Identifying peptides by matching their MS/MS spectra to reference spectra is an effective way of using all spectral information leading to faster and more reliable peptide and protein identifications than current sequence-based methods.

Current peptide identification methods, illustrated below, match each measured MS/MS spectrum against 'theoretical' spectra of all possible peptide sequences. Since relative abundances, neutral losses from parent and product ions, and ratios of products having different charge states are not predictable, this rich, peptide-specific information is not effectively used for establishing identity. Also, prior occurrence information is ignored – each search identifies the peptide as if for the first time.

NIST scientists are performing MS measurements on pure and well-defined mixtures of peptides under carefully controlled conditions. The new library of spectra will provide drug and biomarker discovery scientists with a highly valuable resource against which to benchmark proteomics measurements.

Our approach has been to use collision-induced dissociation (CID) mass spectra of peptide ions are collected in our laboratory with the different mass spectrometers as well as large numbers of spectra are also obtained from other laboratories for analysis of complex clinical specimens. The measurements from our laboratory on pure peptides and well-defined mixtures provide benchmark measurements under carefully controlled conditions. The resulting data from all sources is processed to provide spectra for inclusion into the library: 1) acquire and organize 'Shotgun' proteomics data files from diverse sources, 2) identify peptides with available sequence search engines, 3) create a 'consensus spectrum' from all replicate spectra and find best single spectrum for each peptide ion, 4) derive reliability measures for each spectrum, 5) remove ambiguities and build the library. In addition, as seen above, the same spectrum can have very different scores from different search engines.

To find the consensus spectrum we first align m/z peaks from all spectra, rejects outliers and accept only ions that are present in a majority of the spectra that might have generated the peak. In addition, the best replicate spectrum based on search engine scores and spectrum quality is retained.

We then calculate a measure of reliability by examining a) spectrum/sequence consistency (to match theoretical spectrum, based on relative dissociation rates of adjacent amino acids), b) peptide sequence, and c) peptide class, including those with from semitryptic and other irregular cleavages. The resulting spectra are used to build a library, creating annotated spectra for consensus and best matching single spectra, and resolve problems of similar spectra generating multiple peptides (homologies, small peptides, etc.).

The first version of the peptide library was completed and distributed to several laboratories for testing and the library made available through a public resource site (http:www.proteomecommons.org/tools.asp). It includes data collected in our laboratory and obtained from other laboratories. Automated quality control methods have been developed to deal with the inherent variability of energy, hence spectral patterns, of this class of mass spectra.

All spectra in this library originated from peptide ions generated by electrospray ionization in LC-MS/MS experiments. Most spectra were acquired with ion trap mass spectrometers, with a smaller number from low energy collision cell instruments (qtof-class). Three types of spectra are provided: 1) 'consensus' spectra - derived from multiple (replicate) identifications of a peptide ion, 2) best 'replicate' spectra and 3) high confidence 'single' spectrum identifications.

The library is divided into five sub-libraries:

1) Yeast – 35,135 Consensus, 35,684 Replicate, and 2,462 Single spectra (Saccharomyces Cerevisiae, or Baker's Yeast). This is the most complete collection. It was derived from 2509 data files from 13 laboratories and has served as the principal test-bed for library development.

2) D. Radiodurans – 8,273 Consensus, 8,477 Replicate, and 287 Single spectra – All spectra for this radiation-resistant bacterium were derived from a web-based collection of ion trap spectra provided by PNNL/NCRR [http://ncrr.pnl.gov/data/].

3) M. Smegmatis – 3,569 Consensus, 3,578 Replicate, and 133 Single spectra – All spectra for this bacterium originated from ion trap data files provided in the Open Proteomics Database [http://bioinformatics.icmb.utexas.edu/OPD/].

4) Individual proteins – 3,938 Consensus, 3,922 Replicate, and 15 Single spectra – NIST-measured spectra from separate digests of 19 different proteins on ion trap instruments.

5) Human – 45,377 Consensus, 46,053 Replicate, and 1,940 Single spectra (Homo sapiens) – These spectra were derived from a range of available sources, including labs involved in the Human Proteome Organization plasma proteome project. In addition, it contains spectra from peptides originating from a specific sample types, including leukocytes, hair, and saliva.

This collection is intended primarily to demonstrate the utility of peptide ion fragmentation libraries. Emphasis has been placed on quality control methods, since it has been found that erroneous spectra, especially those containing significant impurity peaks, can lead to false positive results. In order to minimize false negative results (peptide ion not in library), extensive extraction methods were employed. These methods improved the separation of true and false positive results and combined results of several sequence search engines for initial peptide identification. Spectra were annotated to aid spectrum library scoring as well as to document the origin of the spectrum. It is important to note that none of these collections, especially the human library, are complete. Additional measurements will substantially increase both the coverage and the quality of the collection.

Impact: The library is already used for direct peptide identification and for sensitive detection of internal standards, biomarkers, targets proteins with capability to subtract a component from a mixture spectrum. It is sensitive, reliable, fast, and comprehensive, and it can competently search all spectra against library. The library confirms/rejects peptides identified by sequence search programs by comparing to reference spectra. It also links peptides between runs for later processing and identification. It is being incorporated into several proteomics 'pipelines' for integration into practical proteomics analysis.

Future Plans: The libraries, especially those from human samples and standard proteins, are being actively expanded and collaborative work with proteomics centers will increase. We are also providing tested algorithms to search the library and demonstrating its practical applications. The ultimate goal is to have the library installed on and integrated within the data systems of all relevant mass spectrometers and data analysis systems.

Presentations:

- 54th American Society for Mass Spectrometry Conference on Mass Spectrometry (2006).
- Human Proteome Organization (HUPO) 5th Annual World Congress (2006).